Supplementary Materials
Audible Panorama: Automatic Spatial Audio Generation for Panorama Imagery

ACM Reference Format:
. 2019. Supplementary Materials Audible Panorama: Automatic Spatial Audio Generation for Panorama Imagery . In *CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019), May 4–9, 2019, Glasgow, Scotland Uk.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3290605.3300851

1 ADDITIONAL RESULTS

Refer to Figures 1, 2, and 3 for illustrations of the results of running our approach on panorama images.

2 SOUND DATABASE TAGS

Refer to Table 1 for all the audio tags in our database and the number of audio files for each tag.

3 AVERAGE HEIGHTS

Refer to Table 2 for all the average heights of object categories in our approach. Note that these categories match the object tags assigned to objects during object recognition.

4 USER STUDY

Refer to Table 3 for the average ratings given to each sound configuration and scene during the user study. Individual ratings given by all 30 participants can be found in the submitted excel document.

Comments from Participants

All participants considered that the audio assigned by our approach made experiencing the panorama images in VR more immersive. Out of all 30 participants, 26 stated that they thought that we should explore applying our approach to 360°videos.

Some participants also commented that they did not find some of the results believable. Six participants claimed that they did not find the *living room* scene believable because they found that the audio coming from the television would cut off too dramatically when turning their heads away from the sound source.

Some participants commented on the volumes of certain sound sources. In the *Chinatown* scene, for example, some participants commented that the sound of car engines was too loud. In the future we could further explore how to improve the volume of sound sources placed in the scenes.

Transforming Data

Refer to Figure 4 and Table 4 to see the results of performing a log transformation on the results of our user study.

Refer to Table 5 for the results of our post-hoc tests.

In Table 6 we include all of the sets of sound configurations for reference.

CHI 2019, May 4–9, 2019, Glasgow, Scotland Uk

ACM ISBN 978-1-4503-5970-2/19/05…$15.00
https://doi.org/10.1145/3290605.3300851

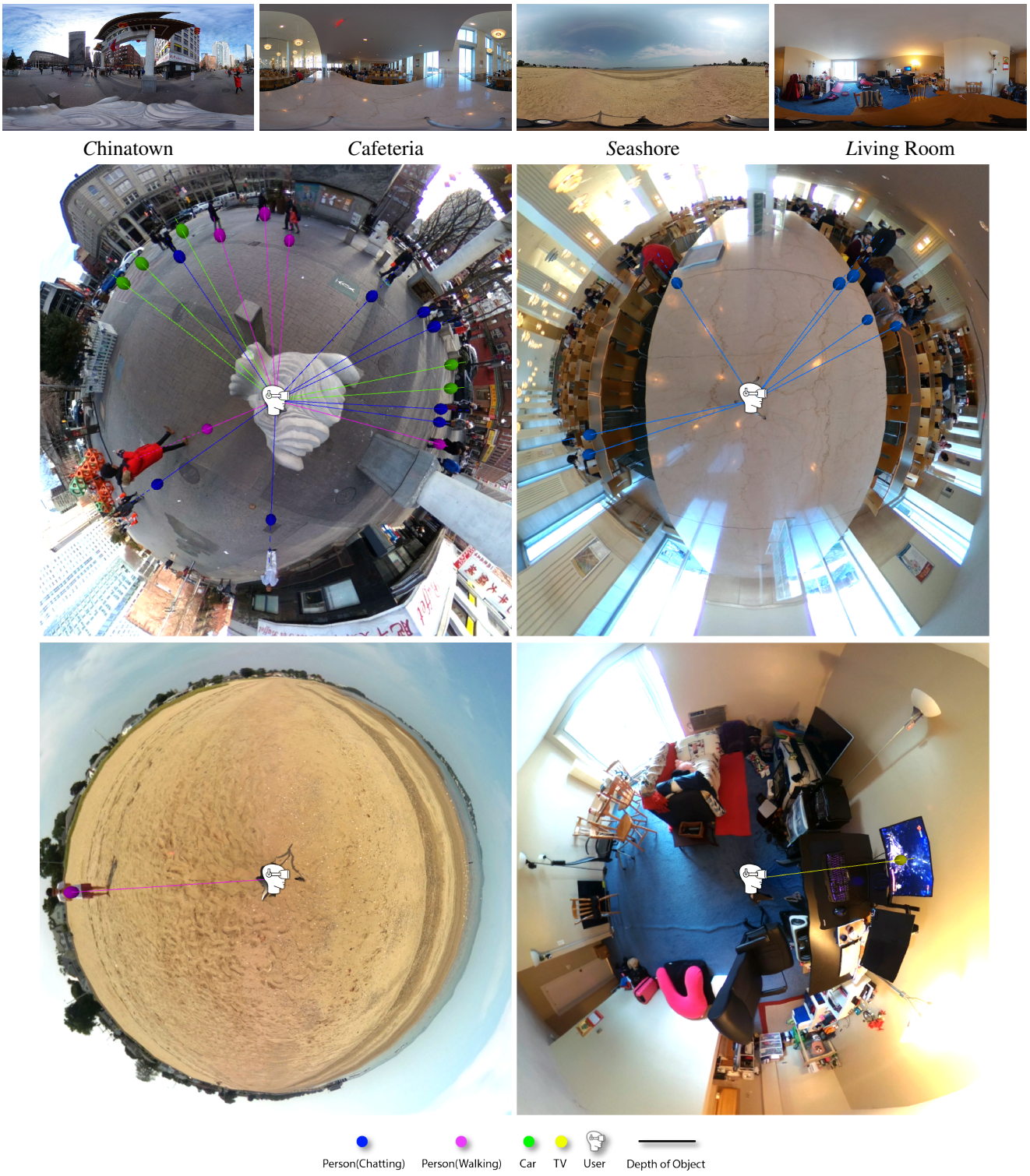

Figure 1: Additional results of running our approach on panorama images. The dots represent the depth at which audio sources are placed for the recognized objects. Audio files that are labeled with audio tags that match the object tags assigned during object recognition are selected to be placed at the audio sources.

Figure 2: Results of running our approach on panorama images. The dots represent the depth at which audio sources are placed for the recognized objects. Audio files that are labeled with audio tags that match the object tags assigned during object recognition are selected to be placed at the audio sources.

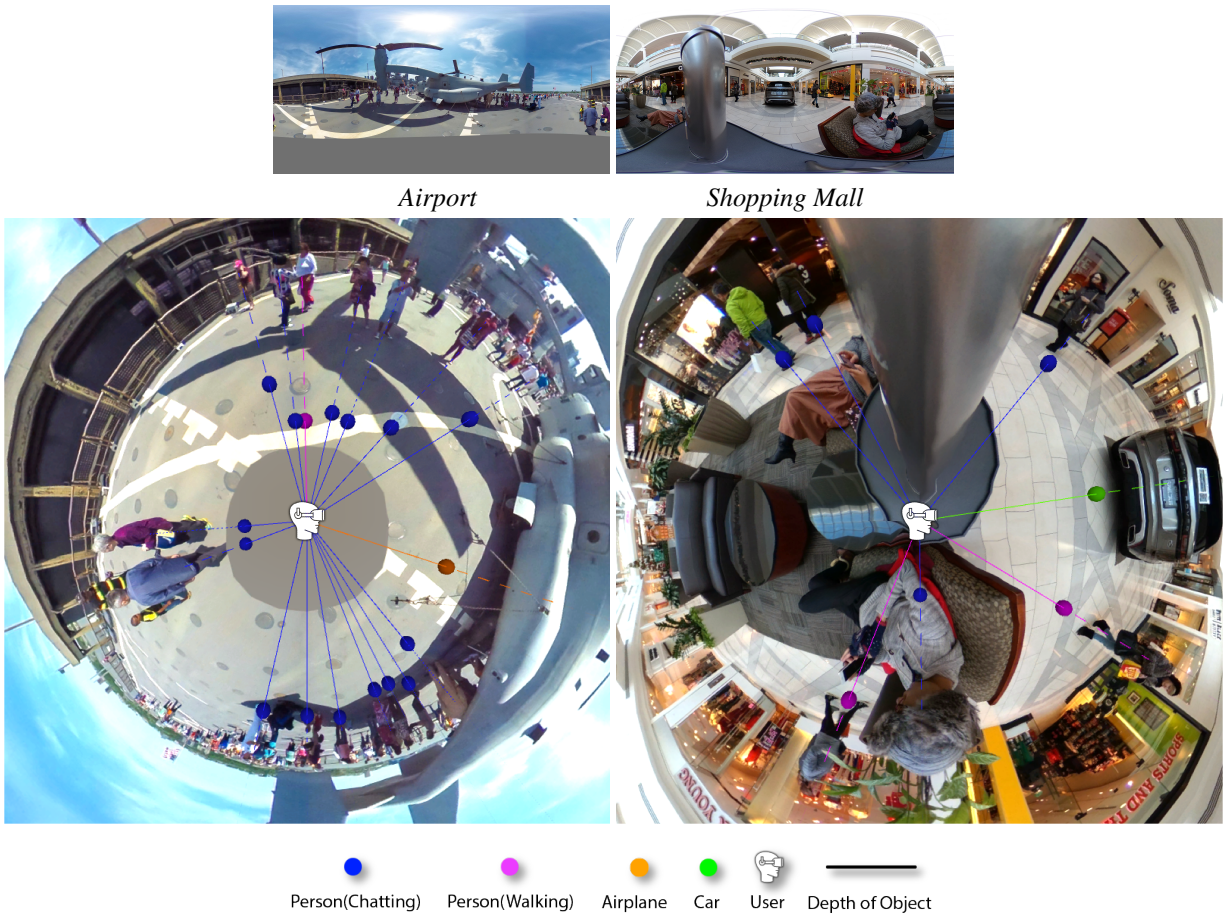

Figure 3: Results of running our approach on panorama images. The dots represent the depth at which audio sources are placed for the recognized objects. Audio files that are labeled with audio tags that match the object tags assigned during object recognition are selected to be placed at the audio sources.

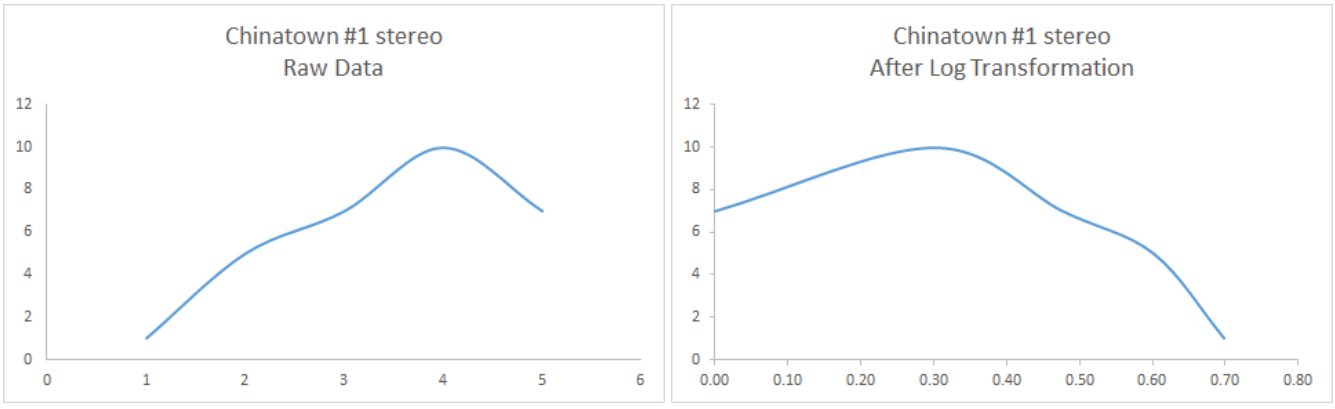

Figure 4: Left: distribution of scores for only stereo audio in the *Chinatown* scene. Right: the same data after applying log transformation with base 10. As can be seen, we do not obtain a normal distribution, which prompts us to further investigate the distribution of our data. Refer to Table 4 for numeric results.

Tag	Type	# of Sounds	Tag	Type	# of Sounds
Audience	Background	12	Airplane	Object	14
Bridge	Background	13	Bicycle	Object	22
City	Background	14	Bird	Object	18
Crowd	Background	12	Boat	Object	12
Escalator	Background	8	Bus	Object	5
Landmark	Background	1	Car	Object	6
Lane	Background	8	Cat	Object	2
Motor Vehicle	Background	8	Cell Phone	Object	8
Piano	Background	9	Chatting	Object	15
Public Space	Background	15	Chatting on Phone	Object	2
Restaurant	Background	5	Clock	Object	21
Road	Background	13	Cow	Object	11
Room	Background	15	Dog	Object	13
Sea	Background	12	Elephant	Object	3
Shopping Mall	Background	9	Horse	Object	13
Snow	Background	10	Laptop	Object	10
Town	Background	9	Motorcycle	Object	15
Transport	Background	7	Refrigerator	Object	17
Water	Background	23	Sheep	Object	12
Waterway	Background	2	Train	Object	21
Wilderness	Background	10	Truck	Object	6
			TV	Object	22
			Typing	Object	12
			Walking	Object	2

Table 1: Sound source files in our database. The tags are used for comparing to the tags given during object recognition in order to select appropriate sound files to be placed. Note that sound source files with both indoor and outdoor versions count as two separate sound source files. The # of sounds shows that how many files of the types in our database.

Tag	Estimated Height in Feet
Airplane	63.0
Bicycle	3.0
Bird	0.3
Boat	10
Bus	10.5
Car	5.0
Cat	0.6
Cell Phone	0.1
Clock	1.2
Cow	4.8
Dog	1.5
Elephant	13.0
Horse	5.0
Laptop	1.2
Motorcycle	3.0
Person*	5.3
Refrigerator	5.5
Sheep	3.5
Train	11.0
Truck	8.0
TV	3.5

Table 2: The average heights for objects used during sound source depth estimation. Note that the objects categories are named using the same object tags used during object recognition. *Person: all object tags matching actions are assumed to be people, so we use average height of persons.

Question #	Configuration Name	Scene			
		Chinatown	Seashore	Cafeteria	Living Room
1	1.spatial	4.07	4.27	3.80	3.33
	1.stereo	3.57	3.90	3.47	2.90
	1.stereo_bg	3.07	3.93	3.13	2.60
2	2.full	3.93	3.90	3.93	3.80
	2.objects	3.83	1.13	3.60	3.80
3	3.real	3.63	3.60	3.20	3.17
	3.synthesized	3.77	3.97	4.03	3.77
4	4.correct	4.17	4.03	3.87	3.60
	4.incorrect	1.00	1.00	1.00	1.00
5	5.EQ-ON	3.70	3.63	4.03	3.90
	5.EQ-OFF	3.83	3.70	3.97	3.87
6	6.depth	4.20	NA	3.73	NA
	6.unidepth	4.03	NA	3.70	NA
	6.rnddepth	3.70	NA	3.47	NA
7	7.10%	3.70	NA	3.30	NA
	7.50%	3.90	NA	3.50	NA
	7.100%	4.07	NA	3.97	NA

Table 3: The average scores given to each scene and sound configuration during our user study. Since the Seashore and Living Room scenes have background audio and only one object sound source in them, the sound configurations for questions 6 and 7 were not tested for them.

Scene	**Set #**						
	1	2	3	4	5	6	7
Chinatown	**0.001**	0.663	0.520	$<$ **0.001**	0.385	0.084	0.281
Cafeteria	**0.034**	0.140	**0.004**	$<$ **0.001**	0.802	0.626	0.119
Seashore	0.242	$<$ **0.001**	0.066	$<$ **0.001**	0.813	NA	NA
Living Room	**0.050**	0.762	0.112	$<$ **0.001**	0.913	NA	NA

Table 4: Results of p-values for each set after performing log transformation on the data with base 10. The results show similar patterns as compared to the original results shown in the main paper. This is further illustrated in Figure 4, where we show the result of applying the log transformation on the distribution of scores for the stereo audio configuration of set 1 for the *Chinatown* scene.

Scene	Set								
	1.ab	1.ac	1.bc	6.ab	6.ac	6.bc	7.ab	7.ac	7.bc
Chinatown	0.057	**<0.001**	0.082	0.450	**0.048**	0.150	0.446	0.155	0.474
Cafeteria	0.210	**0.030**	0.267	0.900	0.312	0.343	0.456	**0.018**	0.097
Seashore	0.071	0.178	0.886	N/A	N/A	N/A	N/A	N/A	N/A
Living Room	0.200	**<0.001**	**0.001**	N/A	N/A	N/A	N/A	N/A	N/A

Table 5: Results of post-hoc tests for Set 1, 6, and 7 for the four scenes. For *Seashore* and *Living Room*, sets 6 (Depth) and 7 (Number of Objects) were not tested since each scene only has the background audio and audio for one object assigned.

Set	Audio Configurations
1.ab	**spatial** & stereo
1.ac	**spatial** & mono
1.bc	stereo & mono
6.ab	**ours** & uniform depth
6.ac	**ours** & random depth
6.bc	uniform & random depth
7.ab	10% & 50% objects
7.ac	10% & **100% objects**
7.bc	50% & **100% objects**

Table 6: A table showing audio configurations descriptions.